We Claim:

1. A fact extraction tool set for extracting information from a document, comprising:

means for annotating a text; and

means for extracting facts from the annotated text.

2. The fact extraction tool set of claim 1, wherein the means for annotating a text comprises means for assigning syntactic and semantic attributes to a text passage by at least one of parsing the text passage and applying text annotation processes other than parsing the text passage.

3. The fact extraction tool set of claim 2, wherein the means for assigning syntactic and semantic attributes to a text passage comprises means for breaking the text passage into its base tokens and annotating the base tokens and patterns of base tokens with a number of orthographic, syntactic, semantic, pragmatic and dictionary-based attributes.

4. The fact extraction tool set of claim 3, wherein the attributes include tokenization, text normalization, part of speech tags, sentence boundaries, parse trees, semantic attribute tagging and other interesting attributes of the text.

5. The fact extraction tool set of claim 2, wherein the means for assigning syntactic and semantic attributes to a text passage comprises independent annotators.

6. The fact extraction tool set of claim 5, wherein the independent annotators use XML as a basis for representing annotated text.

7. The fact extraction tool set of claim 6, further comprising means for resolving conflicting annotation boundaries in the annotated text to produce well-formed XML from the results of independent annotators.

8. The fact extraction tool set of claim 3, wherein the means for breaking the text passage into its base tokens and annotating the base tokens and patterns of base tokens comprises independent annotators, wherein the annotators are of three types comprising:

token attributes, which have a one-per-base-token alignment, where for the attribute type represented, there is an attempt to assign an attribute to each base token;

constituent attributes assigned yes-no values to patterns of base tokens, where the entire pattern is considered to be a single constituent with respect to some annotation value; and

links, which assign common identifiers to coreferring and other related patterns of base tokens.

9. The fact extraction tool set of claim 3, wherein the means for annotating a text further comprises means for associating all annotations assigned to a particular piece of text with the base tokens for that text to generate aligned annotations.

10. The fact extraction tool set of claim 9, wherein the means for extracting facts comprises means for identifying and extracting potentially interesting pieces of information in the aligned annotations by finding patterns in the attributes stored by the annotators.

11. The fact extraction tool set of claim 10, wherein the means for identifying and extracting potentially interesting pieces of information comprises means for recognizing both true left and right constituent attributes and non-contiguous constituent attributes.

12. The fact extraction tool set of claim 10, wherein the means for identifying and extracting potentially interesting pieces of information comprises at least one text pattern recognition rule written in a rule-based information extraction language, wherein the at least one text pattern recognition rule queries for at least one of literal text, attributes, and relationships found in the aligned annotations to define the facts to be extracted.

13. The fact extraction tool set of claim 12, wherein the at least one text pattern recognition rule can use regular expression functionality, XPath-based functionality, and auxiliary definitions in any combination.

14. The fact extraction tool set of claim 12, wherein the at least one text pattern recognition rule comprises a pattern that describes the text of interest, a label that names the pattern for testing and debugging purposes; and an action that indicates what should be done in response to a successful match.

15. The fact extraction tool set of claim 12, wherein the means for identifying and extracting potentially interesting pieces of information further comprises at least one auxiliary definition statement used to name and define a fragment of a pattern.

16. A rule-based information extraction language for use in identifying and extracting potentially interesting pieces of information in aligned annotations in a text, comprising at least one text pattern recognition rule that queries for at least one of literal text, attributes, and relationships found in the aligned annotations to define the facts to be extracted.

17. The language of claim 16, wherein the at least one text pattern recognition rule can use regular expression functionality, XPath-based functionality, and auxiliary definitions in any combination.

18. The language of claim 16, wherein the at least one text pattern recognition rule comprises a pattern that describes the text of interest, a label that names the pattern for testing and debugging purposes, and an action that indicates what should be done in response to a successful match.

19. The language of claim 16, further comprising at least one auxiliary definition statement used to name and define a fragment of a pattern.

20. A text annotation tool comprising:

means for assigning syntactic and semantic attributes to a text passage by at least one of parsing the text passage and applying text annotation processes other than parsing the text passage, including means for breaking the text passage into its base tokens and annotating the base tokens and patterns of base tokens with a number of orthographic, syntactic, semantic, pragmatic and dictionary-based attributes; and

means for associating all annotations assigned to a particular piece of text with the base tokens for that text to generate aligned annotations.

21. The text annotation tool of claim 20, wherein the attributes include tokenization, text normalization, part of speech tags, sentence boundaries, parse trees, semantic attribute tagging and other interesting attributes of the text.

22. The text annotation tool of claim 20, wherein the means for assigning syntactic and semantic attributes to a text passage comprises independent annotators.

23. The text annotation tool of claim 22, wherein the independent annotators use XML as a basis for representing annotated text.

24. The text annotation tool of claim 23, further comprising means for resolving conflicting annotation boundaries in the annotated text to produce well-formed XML from the results of independent annotators.

25. The text annotation tool of claim 20, wherein the means for breaking the text passage into its base tokens and annotating the base tokens and patterns of base tokens comprises independent annotators, wherein the annotators are of three types comprising:

token attributes, which have a one-per-base-token alignment, where for the attribute type represented, there is an attempt to assign an attribute to each base token;

constituent attributes assigned yes-no values to patterns of base tokens, where the entire pattern is considered to be a single constituent with respect to some annotation value; and

links, which assign common identifiers to coreferring and other related patterns of base tokens.

26. A computer program product for extracting information from a document, the computer program product comprising a computer usable storage medium having computer readable program code means embodied in the medium, the computer readable program code means comprising:

computer readable program code means for annotating a text; and

computer readable program code means for extracting facts from the annotated text.

27. The computer program product of claim 26, wherein the computer readable program code means for annotating a text comprises computer readable program code means for assigning syntactic and semantic attributes to a text passage by at least one of parsing the text passage and applying text annotation processes other than parsing the text passage.

28. The computer program product of claim 27, wherein the computer readable program code means for assigning syntactic and semantic attributes to a text passage comprises computer readable program code means for breaking the text passage into its base tokens and annotating the base tokens and patterns of base tokens with a number of orthographic, syntactic, semantic, pragmatic and dictionary-based attributes.

29. The computer program product of claim 28, wherein the attributes include tokenization, text normalization, part of speech tags, sentence boundaries, parse trees, semantic attribute tagging and other interesting attributes of the text.

30. The computer program product of claim 27, wherein the computer readable program code means for assigning syntactic and semantic attributes to a text passage comprises independent annotators.

31. The computer program product of claim 30, wherein the independent annotators use XML as a basis for representing annotated text.

32. The computer program product of claim 31, further comprising computer readable program code means for resolving conflicting annotation boundaries in the annotated text to produce well-formed XML from the results of independent annotators.

33. The computer program product of claim 28, wherein the computer readable program code means for breaking the text passage into its base tokens and annotating the base tokens and patterns of base tokens comprises individual annotators, wherein the annotators are of three types comprising:

token attributes, which have a one-per-base-token alignment, where for the attribute type represented, there is an attempt to assign an attribute to each base token;

constituent attributes assigned yes-no values to patterns of base tokens, where the entire pattern is considered to be a single constituent with respect to some annotation value; and

links, which assign common identifiers to coreferring and other related patterns of base tokens.

34. The computer program product of claim 28, wherein the computer readable program code means for annotating a text further comprises computer readable program code means for associating all annotations assigned to a particular piece of text with the base tokens for that text to generate aligned annotations.

35. The computer program product of claim 34, wherein the computer readable program code means for extracting facts comprises computer readable program code means

for identifying and extracting potentially interesting pieces of information in the aligned annotations by finding patterns in the attributes stored by the annotators.

36. The computer program product of claim 35, wherein the computer readable program code means for identifying and extracting potentially interesting pieces of information further comprises computer readable program code means for recognizing both true left and right constituent attributes and non-contiguous constituent attributes.

37. The computer program product of claim 35, wherein the computer readable program code means for identifying and extracting potentially interesting pieces of information comprises at least one text pattern recognition rule written in a rule-based information extraction language, wherein the at least one text pattern recognition rule queries for at least one of literal text, attributes, and relationships found in the aligned annotations to define the facts to be extracted.

38. The computer program product of claim 37, wherein the at least one text pattern recognition rule can use regular expression functionality, XPath-based functionality, and auxiliary definitions in any combination.

39. The computer program product of claim 37, wherein the at least one text pattern recognition rule comprises a pattern that describes the text of interest, a label that names the pattern for testing and debugging purposes, and an action that indicates what should be done in response to a successful match.

40. The computer program product of claim 37, wherein the computer readable program code means for identifying and extracting potentially interesting pieces of information further comprises at least one auxiliary definition statement used to name and define a fragment of a pattern.

41. A method of extracting information from a document, comprising the steps of:
annotating a text; and
extracting facts from the annotated text.

42. The method of claim 41, wherein the step of annotating a text comprises assigning syntactic and semantic attributes to a text passage by at least one of parsing the text passage and applying text annotation processes other than parsing the text passage.

43. The method of claim 42, wherein the parsing of the text passage comprises breaking it into its base tokens and annotating the base tokens and patterns of base tokens with a number of orthographic, syntactic, semantic, pragmatic and dictionary-based attributes.

44. The method of claim 43, wherein the attributes include tokenization, text normalization, part of speech tags, sentence boundaries, parse trees, semantic attribute tagging and other interesting attributes of the text.

45. The method of claim 42, wherein the parsing of the text passage is carried out by independent annotators.

46. The method of claim 45, wherein the individual annotators use XML as a basis for representing annotated text.

47. The method of claim 46, further comprising the step of resolving conflicting annotation boundaries in the annotated text to produce well-formed XML from the results of independent annotators.

48. The method of claim 43, wherein the step of breaking the text passage into its base tokens and annotating the base tokens and patterns of base tokens is carried out using independent annotators, wherein the annotators are of three types comprising:

token attributes, which have a one-per-base-token alignment, where for the attribute type represented, there is an attempt to assign an attribute to each base token;

constituent attributes assigned yes-no values to patterns of base tokens, where the entire pattern is considered to be a single constituent with respect to some annotation value; and

links, which assign common identifiers to coreferring and other related patterns of base tokens.

49. The method of claim 43, wherein the step of annotating a text further comprises the step of associating all annotations assigned to a particular piece of text with the base tokens for that text to generate aligned annotations.

50. The method of claim 49, wherein the step of extracting facts comprises identifying and extracting potentially interesting pieces of information in the aligned annotations by finding patterns in the attributes stored by the annotators.

51. The method of claim 50, wherein the step of identifying and extracting potentially interesting pieces of information comprises recognizing both true left and right constituent attributes and non-contiguous constituent attributes.

52. The method of claim 50, wherein the patterns are found using at least one text pattern recognition rule written in a rule-based information extraction language, wherein the at least one text pattern recognition rule queries for at least one of literal text, attributes, and relationships found in the aligned annotations to define the facts to be extracted.

53. The method of claim 52, wherein the at least one text pattern recognition rule can use regular expression functionality, XPath-based functionality, and auxiliary definitions in any combination.

54. The method of claim 52, wherein the at least one text pattern recognition rule describes the text of interest, names the pattern for testing and debugging purposes; and indicates what should be done in response to a successful match.

55. The method of claim 52, wherein the patterns are found further using at least one auxiliary definition statement used to name and define a fragment of a pattern.